A Reconstruction Method for Cross-Cut Shredded Documents Based on the Extreme Learning Machine Algorithm

Zhenghui Zhang (344237028@qq.com)
Xiangtan University
Juan Zou
Xiangtan University
Jinhua Zheng
Xiangtan University
Shengxiang Yang
Xiangtan University
Dunwei Gong
Xiangtan University
Tingrui Pei
Xiangtan University

Research Article

Keywords: Reconstruction of cross-cut shredded text documents (RCCSTD), extreme learning machine algorithm, consensus information, pairwise compatibility measurement model

Posted Date: January 4th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1143560/v1

A reconstruction method for cross-cut shredded documents based on the extreme learning machine algorithm

Zhenghui Zhang [a,b,1], Juan Zou[a,b,1,*], Shengxiang Yang[a,d], Jinhua Zheng[a,c], Dunwei Gong[a,e], Tingrui Pei[a,b]

[a]*Key Laboratory of Hunan Province for Internet of Things and Information Security, Xiangtan University, Xiangtan, Hunan, 411105, China*
[b]*Key Laboratory of Intelligent Computing and Information Processing, Ministry of Education, Information Engineering College of Xiangtan University, Xiangtan, Hunan, 411105, China.*
[c]*Hunan Provincial Key Laboratory of Intelligent Information Processing and Application, Hengyang, 421002, China.*
[d]*School of Computer Science and Informatics, De Montfort University, Leicester LE1 9BH,U.K.*
[e]*School of Information and Electronic Engineering, China University of Mining and Technology, Xuzhou, 221116, China.*

Abstract

Reconstruction of cross-cut shredded text documents (RCCSTD) has important applications for information security and judicial evidence collection. The traditional method of manual construction is a very time-consuming task, so the use of computer-assisted efficient reconstruction is a crucial research topic. Fragment consensus information extraction and fragment pair compatibility measurement are two fundamental processes in RCCSTD. Due to the limitations of the existing classical methods of these two steps, only documents with specific structures or characteristics can be spliced, and pairing error is larger when the cutting is more fine-grained. In order to reconstruct the fragments more effectively, this paper improves the extraction method for consensus information and constructs a new global

*Corresponding author: Juan Zou
Email address: zoujuan@xtu.edu.cn (Juan Zou)

1 The first author and the second author have the same contribution and should be considered as a co-first author.

pairwise compatibility measurement model based on the extreme learning machine algorithm. The purpose of the algorithm's design is to exploit all available information and computationally suggest matches to increase the algorithm's ability to discriminate between data in various complex situations, then find the best neighbor of each fragment for splicing according to pairwise compatibility. The overall performance of our approach in several practical experiments is illustrated. The results indicate that the matching accuracy of the proposed algorithm is better than that of the previously published classical algorithms and still ensures a higher matching accuracy in the noisy datasets, which can provide a feasible method for RCCSTD intelligent systems in real scenarios.

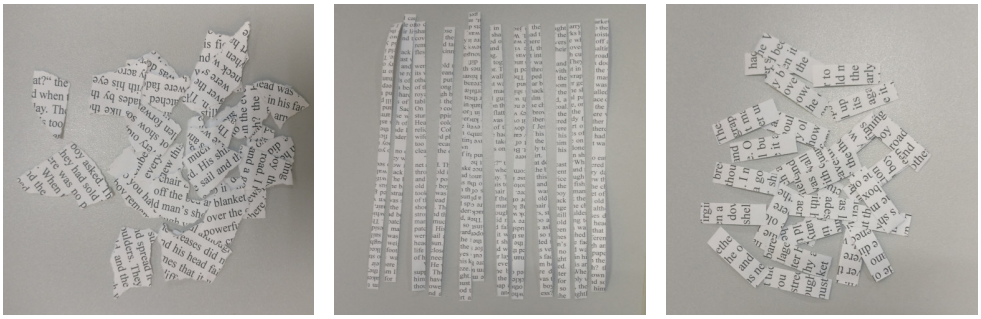
Keywords: Reconstruction of cross-cut shredded text documents (RCCSTD), extreme learning machine algorithm, consensus information, pairwise compatibility measurement model

1. Introduction

Shredding sensitive documents is a common way to destroy evidence. Therefore, reconstruction of shredded paper is a crucial task in the field of forensic evidence. It has also been successfully applied in the fields of military intelligence acquisition and archaeology. The structure or characteristics of the fragments after documents are shredded because the edge information of the pixels in the text is reduced. The conventional manual reconstruction method is difficult and time-consuming Heingartner (2003). Therefore, efficient and accurate automatic computer-assisted reconstruction is desired.

Reconstructing shredded documents requires a large number of fragments to be spliced together without any prior information. For example, in some scenarios, the words and symbols in the fragments are unknown, and it is unknown how many rows and columns the document has been cut into. In the field of document reconstruction, different reconstruction methods are used for different types of shredding. Document reconstruction has been divided into three kinds of related problems according to the following three shredding methods. The first is hand-torn documents (HTD) which have irregular boundaries that have unique geometric texture information (see Fig. 1(a)) that can be matched by local geometric texture curves Zhu et al. (2007). The second is the strip shredded text documents (SSTD), which are evenly cut by a shredder and produce the strip fragment (see Fig. 1(b)). Existing research

has obtained good reconstruction results with strip shredded documents Lin & Fan-Chiang (2012); Ukovich et al. (2004). The third is the cross-cut shredded text documents (CCSTD), which are produced by the shredder making even horizontal and vertical cuts. Due to the fragments being cut smaller, it is difficult to obtain useful edge pixel feature information (see Fig. 1(c)). This paper mainly focuses on finding a feasible solution for the Reconstruction of cross-cut shredded text documents (RCCSTD) Biesinger et al.; Gong et al. (2016); Chen et al. (2019) problem.



(a) Hand-torn documents (b) Strip shredded text documents (c) Cross-cut shredded documents

Figure 1: Three different methods of shredding

Early research has made good progress on the reconstruction of strip shredded text documents (RSSTD) Ukovich et al. (2004); Prandtstetter & Raidl (2008); Lin & Fan-Chiang (2012), but these methods cannot reconstruct the document when the cutting is more fine-grained. Prandtstetter & Raidl (2008) proposed a novel approach to RSSTD. In this approach, the combination problem of fragments is defined as a cost function, and the traveling salesman problem (TSP) is used to find the least cost combination to obtain the correct splicing. Based on this work, many researchers have transformed the RCCSTD problem into a TSP.

Wang & Ji (2014) proposed a two-step strategy. The first step classifies the fragments of the same row by using a clustering algorithm to narrow the scope of the paired search. The second step reconstructs the rows of the document and then stitches the rows to restore the document. This is a viable solution, but it is difficult to guarantee the accuracy of the pairing. Xu et al. (2014) presented a feature-matching algorithm based on character recognition using a database of the letters. Like the two-step strategy,

this method reconstructs the shredded document by row clustering, intra-row splicing, and inter-row splicing. This method improves the matching accuracy by classifying the fragments in different rows and reducing the range of matching search. Chen et al. (2019) improved the clustering method by using horizontal projection and the constrained seed K-means algorithm. However, this method is dependent on the row features of the text in each fragment, and you must know the first fragment of each row as the initialization center of the cluster, which leads to limitations in the algorithm. For example, when a document is cut into more fragments, the number of fragments with white edges will increase, and it is impossible to know how many rows the document was cut into. In this case, the clustering-based splicing method is disadvantageous. Different from the two-step strategy, Gong et al. (2016) proposed a two-dimensional TSP reconstruction method, which defines a comprehensive cost function for global optimization. However, when the number of fragments increases, there is less pixel information available at the edge of the fragments. The comprehensive cost function is prone to uncontrollable degradation, which leads to the deterioration of the accuracy of paired matching. Sleit et al. (2013) defined a new cost function, which mainly relies on black pixels to measure the cost of pairing two fragments together. However, this method is sensitive to cost function in complex datasets, resulting in low splicing accuracy. Paixão et al. (2019), while exploring character shapes as visual features for compatibility computation, proposed a training scheme based on digitally simulated document shredding from a well-known OCR database. However, this method cannot complete the restoration of documents without prior information. Many examples in real life, such as patterns or symbols in cultural relics, do not have prior information.

There are three methods for solving the RCCSTD problem as follows:

1) Two-stage: In the first step, the fragments of different rows are classified by the clustering algorithm. Then each row is reconstructed by the TSP problem. The second step uses feature matching to merge the reconstructed fragments. This is currently the most commonly used method to solve the RCCSTD Wang & Ji (2014); Xu et al. (2014); Chen et al. (2019). However, because the clustering algorithm relies too much on the row distribution features, the scope of application is limited, and the error of the cluster will further affect the accuracy of the row reconstruction so that the global error increases.

2) Comprehensive cost function: This method considers RCCSTD to be

a two-dimensional TSP problem that uses the idea based on the square jigsaw puzzle approach Freeman & Garder (1964) to find the optimal combination problem through the comprehensive global cost function Gong et al. (2016). Since RCCSTD is significantly different from typical puzzle problems, the splicing accuracy of the algorithm decreases rapidly with an increase in the number of fragments, and it is unable to deal with the situation of missing fragments.

3) Matching cost: A cost function is defined to measure the cost of pairing two fragments together Sleit et al. (2013); Ranca (2013). Only the edge pixel features are considered, and the interior features of fragments are not fully used. This kind of method is sensitive to the cost function and cannot be used in scenes when the cutting is more fine-grained.

Fragment consensus information extraction and fragment pair compatibility measurement are two fundamental processes in RCCSTD solutions. Existing works adopt some general methods in these two processes, but there are limitations in their use since they ignore the specific structure or characteristics of the RCCSTD. Therefore, in this paper, we propose a new consensus information extraction method and construct a new global pairwise compatibility measurement model based on the extreme learning machine algorithm. The structure of the RCCSTD problem in real scenarios is considered comprehensively. The comprehensive consensus information includes the edge pixel feature of the fragment and the text distribution feature inside the fragment, which increases the algorithm's ability to discriminate between the unmatched fragments. A single hidden layer feedforward neural network is used to predict the expected matching edge of the fragment and to evaluate the overall difference from the fragment to be matched. When each element in a pair independently thinks that the other element is its most likely match, it prefers matching in pairs, which is called "Best Buddies" Dekel et al. (2015). The advantage of using neural networks in compatibility measurement is that by randomly mapping the one-dimensional discrete features of debris into high-dimensional space, the independence of sample features is guaranteed, and the dispersion degree of data increases, so that the data have a higher discrimination degree on the projected dimension. Extreme learning machines are feedforward neural networks for feature learning with a single layer or multiple layers of hidden nodes (Huang et al. (2004)). The weights from the input layer to the hidden layer are randomly determined once and do not need to be adjusted during the training. The weights

from the hidden layer to the output layer only need to be determined by solving a linear simultaneous equation, which can converge rapidly on a few samples.

In a real scene, the shredder is prone to producing uncontrollable edge noise in the process of cutting, which makes the algorithm's requirements during the splicing process more demanding. In some individual cases, manual splicing still makes it difficult to avoid the occurrence of errors, and it is difficult to distinguish between the correct match (see Fig. 2(a)) and the wrong match (see Fig. 2(b)). When the document is cut into more fragments, this special case can more easily happen. Different from manual splicing, computer-aided semi-auto splicing usually regards splicing work as the optimal problem of finding a combination of fragments. However, the previous methods usually do not consider the fault-tolerant mechanism of splicing. When a pairing error occurs, the wrong pairing can only be corrected by manual search. In the splicing based on the TSP strategy, even a single mismatch can cause a global mismatch, and the superposition of errors leads to an increase in the number of manual interventions, which reduces splicing efficiency. In order to solve this complex situation, our model establishes a candidate set for each erroneous fragment. This effectively assists in the manual correction of error splicing and analyzes the effectiveness of the candidate strategy through experiments.

The main novelties of this paper are as follows:

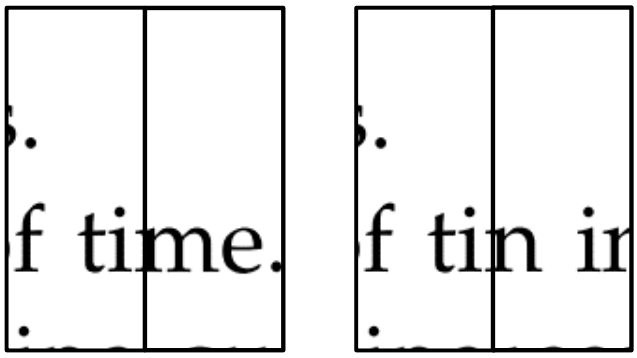
1) Comprehensive consensus information is used to improve the ability of the algorithm to discriminate between the data. At the same time, the edge pixel information and the edge distribution information of the text are considered, and all available information is used for suggestion matching so that the algorithm can be applied to the restoration of all types of documents.

2) A pairing compatibility measurement model based on the extreme learning machine algorithm is proposed. This model improves pairing accuracy and solves the problem of too much sensitivity to abnormal edge pixels in compatibility measurement.

3) The greedy strategy is used to combine all the pairing information and get the correct line sequence by continuously looping and adjusting.

4) In the process of manual error correction, candidate matching sets are

used to improve the efficiency of error correction and greatly reduce the amount of manual feedback required.



(a) (b)

Figure 2: A special case: It is difficult to distinguish the splicing right (a) or wrong (b)

The remainder of the paper is organized as follows. The formal model of RCCSTD is defined in Section 2 and a detailed solution is given. Section 3 provides and discusses the experimental results in detail. Finally, this paper is concluded in Section 4.

2. Methodology

In this section, we describe in detail the definition of the RCCSTD problem, the specific extraction method of the consensus information, and how to use the consensus information to calculate the matching compatibility score.

2.1. Definition of problem

The definition of the RCSSTD problem is similar to the definition of the jigsaw puzzle approaches Chen et al. (2019). Gong et al. (2016) present a formal problem description of RCCSTD as a combinatorial optimization problem based on Prandtstetter (2009). Given that the shape and direction of the paper fragments are known, a given number of document fragments is $(N \times M)$; each fragment has a unique number, and the number of all fragments can be defined as a two-dimensional matrix S. The location of

the fragments in the matrix S is the only unknown. The purpose of the algorithm is to construct a matrix corresponding to the original document.

The RGB color model is a common form of image storage in computers, in which red, green and blue light are added together. Given that the document is usually composed of white background and black text, in order to avoid unnecessary calculation, it is common to convert the shredded picture into the form of grayscale. We obtain image edge features by extracting image edge pixels. For each fragment, S_l^i and S_r^i are used to represent the left edge and right edge pixel vectors of the ith fragment, respectively. The mathematical representation is as follows:

$$\begin{aligned} S_l^i &= \left[x_1^l, x_2^l, x_3^l, \cdots, x_h^l\right], \quad x_i^l \in [0, 255], 1 \leq i \geq h, \\ S_r^i &= \left[x_1^r, x_2^r, x_3^r, \cdots, x_h^r\right], \quad x_i^l \in [0, 255], 1 \leq i \geq h, \end{aligned} \tag{1}$$

where h is the edge length of the picture.

2.2. Image enhancement (IE) and consensus information

The number and distribution of black pixels at the edges is crucial to the splicing results in the process of splicing fragments. During the reconstructing process, we analyzed the error cases, which can be summarized as follows:

1) When the text or symbol in the document is cut, and the position of cutting is unpredictable, the pixels of the edges corresponding to the correctly matched fragments are not one-to-one (see Fig. 3). If the effective pixels of the edge used for pairing are not enough to reflect the compatibility of pairing, a splicing error can easily occur.

2) Because algorithms usually only rely on the edge pixel features of the fragment to splice, it is easy to ignore the useful information provided by the non-edge features.

In order to deal with this situation, two methods are used to extract consensus information, one is to divide the pixel and extend the edge features, the other is to perform Gaussian filtering on the image to obtain the distribution characteristics of the text at the edge. We shall discuss this in more detail later. In our experiments, we will verified the effectiveness of these two strategies.

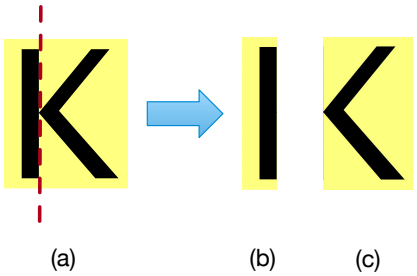


Figure 3: When the character "K" is randomly cut, the adjacent edge pixels of the fragment being cut are not one-to-one.

2.2.1. Pixel division

As the document is cut into more fragments, the structure of the edges becomes more complex, and the edge of the fragment represents a white pixel increase. In order to deal with these special cases, we propose a "cell splitting" method to increase the information distribution of the edge pixels. The edge matching of text mainly depends on the compatibility of black pixels. Improving neighbor consensus information can improve matching accuracy. The process can be defined as:

$$
\begin{aligned}
S &= [x_1, x_2, x_3, \cdots, x_h], \quad x_i \in [0, 255], \quad 1 \leq i \leq h, \\
S^{\dagger} &= \left[\underbrace{x_1, \cdots, x_1}_{m}, \underbrace{x_2, \cdots, x_2}_{m}, \cdots, \underbrace{x_h, \cdots, x_h}_{m}, \right], \quad x_i \in [0, 255], 1 \leq i \leq h,
\end{aligned}
\tag{2}
$$

where h is the edge length of the picture, and m is the number of pixel divisions, which is set as 3 in our experiments.

2.2.2. Text distribution feature

In the previous splicing strategy Chen et al. (2019); Xu et al. (2014), it is difficult to obtain correct pairing in a large number of fragments by only considering the edges of the pixels. Therefore, the row features of the text in the fragments are used as important information to classify different rows. Gaussian smoothing is adopted to preprocess the image so that the effective information in the fragment is focused on the text in the fragment, and the influence of the text information on the surrounding pixels is amplified. That is, the color represented by the pixel closest to the text will be deepened,

which enriches the useful information of the fragment. Fig. 4 shows the effect of Gaussian filtering on the edge.

$$S^* = [x_1^*, x_2^*, x_3^*, \cdots, x_h^*], \quad x_i \in [0, 255], \quad 1 \leq i \leq h, \tag{3}$$

where h is the edge length of the picture and x_h^* is the edge pixel of the processed fragment.

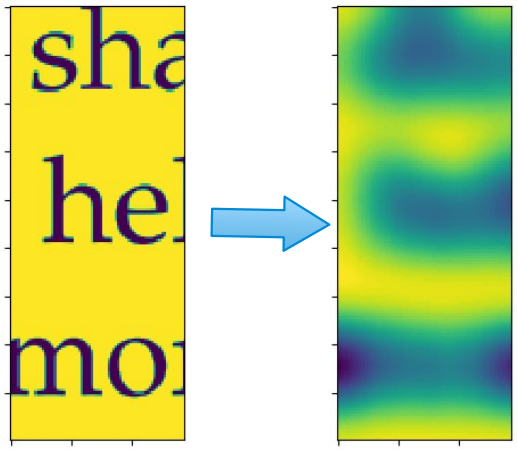


Figure 4: Gaussian convolution is used to enhance edge information

2.2.3. Extracting consensus information

After completing the steps in Section 2.2.1 and Section2.2.2, all the information needs to be combined to get the final consensus information. The final consensus information vector is defined as follows:

$$X = \begin{cases} [S^\dagger, S^*] & \text{Plain text dataset with row features,} \\ [S^\dagger] & \text{Otherwise.} \end{cases} \tag{4}$$

2.3. Pairwise-compatibility measurement model based on the extreme learning machine algorithm

In the previous section, we completed the work of obtaining image edge pixel consensus information. We can now propose a pairwise-compatibility measurement model (PCMM), which is an extremely fast learning algorithm for the single hidden layer feedforward neural network based on the extreme learning machine algorithm. The edge pixels of the fragments to be matched are used to train the neural network (see Fig. 5). In order to unify data

standards and improve the comparability of the data, all samples and labels need to be normalized by z-score Jain et al. (2005). the model is divided into the following steps:

Step 1: Obtain the consensus information matrix X at the right edge of the fragments to be matched by the method in Section 2.2.3 to serve as the input of the neural network. The mathematical model of the standard single hidden layer feedforward neural network Jain et al. (1996) can be defined as:

$$\sum_{i=1}^{\tilde{N}} \beta_i g\left(\mathbf{w}_i \cdot \mathbf{x}_j + b_i\right) = \mathbf{o}_k, \quad j = 1, \cdots, N, \quad k = 1, \cdots, h \tag{5}$$

where N is the number of neurons in the input layer; $\tilde{N}$ is the number of neurons in the hidden layer; h is the edge length of the picture, and $N > h$; $\mathbf{w}_i$ is the weight vector connecting the ith hidden neuron and the input neurons; β_i is the weight vector connecting the ith hidden neuron and the output neurons, and b_i is the threshold of the ith hidden neuron; $\mathbf{w}_i \cdot \mathbf{x}_j$ denotes the inner product of $\mathbf{w}_i$ and $\mathbf{x}_j$. And $g(x)$ represents an activation function.

The above $\tilde{N}$ equations can be formulated compactly as

$$\mathbf{H}\beta = \mathbf{T} \tag{6}$$

H is called the hidden output matrix of the neural network, where

$$\begin{aligned} &\mathbf{H}\left(\mathbf{w}_1, \cdots, \mathbf{w}_{\tilde{N}}, b_1, \cdots, b_{\tilde{N}}, \mathbf{x}_1, \cdots, \mathbf{x}_N\right) \\ &= \begin{bmatrix} g\left(\mathbf{w}_1 \cdot \mathbf{x}_1 + b_1\right) & \cdots & g\left(\mathbf{w}_{\tilde{N}} \cdot \mathbf{x}_1 + b_{\tilde{N}}\right) \\ \vdots & \cdots & \vdots \\ g\left(\mathbf{w}_1 \cdot \mathbf{x}_N + b_1\right) & \cdots & g\left(\mathbf{w}_{\tilde{N}} \cdot \mathbf{x}_N + b_{\tilde{N}}\right) \end{bmatrix}_{N \times \tilde{N}, N \neq \tilde{N}} \end{aligned} \tag{7}$$

The output weight β is the smallest norm leastsquares solution according to the least square theory:

$$\hat{\beta} = \mathbf{H}^{\dagger}\mathbf{T} \tag{8}$$

where $\mathbf{H}^{\dagger}$ is represented as the Moore-Penrose generalized inverse of H (Li et al. (2021)), and $H^{\dagger} = \left(H^T H\right)^{-1} H^T$.

Step 2: Randomly initialize the input layer weights w_i and bias b_i.

Step 3: The edge pixel matrix of the fragments to be matched is normalized by z-score and treated as the expected output vector T of the neural

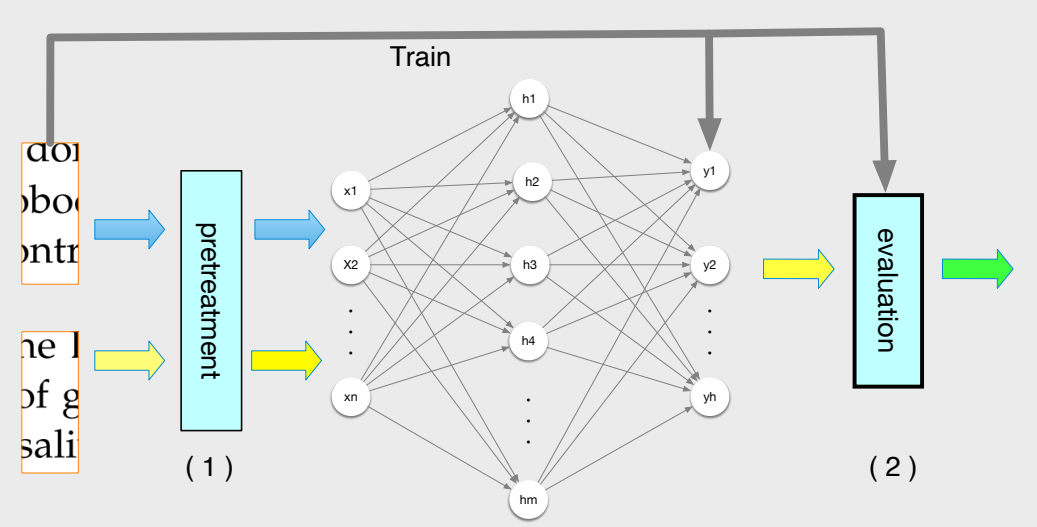


Figure 5: A self-supervised feature matching model for cross-cut shredded documents. (1) The consensus information is extracted and normalized by z-score. (2) The degree of difference between the prediction feature and matching feature is analyzed.

network, which is put into formula 9 to train to get the weight vector β^* of the output layer.

$$\beta^* = \left(H^T H + \frac{1}{C}\right)^{-1} H^T T, \tag{9}$$

where H is called the hidden layer output matrix of the neural network Huang & Babri (1998); C is the regularization coefficient, which is considered $1e5$ in the experiment.

Step 4: The consensus information vector of the left edge of each fragment is taken as the input of the PCMM. The predicted feature vector of the candidate fragment is obtained according to vector β^* in formula 5. The mean square error (MSE) Murphy (1988) is used to measure the overall error between the predicted feature vector and the edge pixel feature vector to be matched. The compatibility score between the matching fragment and each candidate matching fragment is obtained.

Step 5: The candidate sets in Step4 are sorted by pairwise compatibility measurement score. The candidate fragments with the smallest candidate concentration errors are selected as the best match.

After these five steps, the best candidate matching and corresponding candidate set of each scrap can be quickly obtained, and the computer com-

pletes these processes automatically.

2.4. Semi-automatic greedy reconstruction

2.4.1. Greedy reconstruction

The pairing compatibility metric in Section 2.3 is used to construct the candidate neighbor set corresponding to each fragment. Considering that the relationship between pairing information is independent of each fragment, to completely reconstruct the document, all the pairing information must be combined. However, due to potential errors in pairing, it is not possible to directly reconstruct the document, and it is necessary to adjust the wrong pairings constantly. Previous methods required a manual search to complete the error-matching correction process, but this was an inefficient approach. We are inspired by the puzzle algorithm proposed by Paikin & Tal (2015). The process of constructing the pairing information and marking the incorrect matches is repeated until no error matches exist. The algorithm proposed in this paper has high matching accuracy, so only a small amount of work is needed to refine it. Reconstruction based on the greedy strategy happens through the following steps:

Step 1: Start by randomly selecting a pair from the pairwise set, and combine the pairs that can be matched according to the pairwise information. Step 2: Repeat step 1 until all spliced sequences in the collection are unchanged.

Step 3: According to the splicing result of Step 2, the error pair is manually marked. Go to step two until there are no errors in the results.

Through these three steps, the rows of the document are completely reconstructed. The splicing between rows is relatively easy. Because the inter-row splicing conforms to the characteristics of the RSSTD problem, it can be transformed into a RSSTD problem for reconstruction.

2.4.2. Feedback-based splicing

In the process of automatic splicing, due to some unpredictable factors (See Fig. 2) the pairwise compatibility measurement cannot match all fragments to true neighbors. In the classical algorithms, it is necessary to manually adjust the mismatched fragments after the computer splicing is completed. In order to reduce the cost of manual splicing, we construct a candidate set of neighbors for each fragment, and in the case of a match error, use it to assist in manual correction. Because the number of mismatched

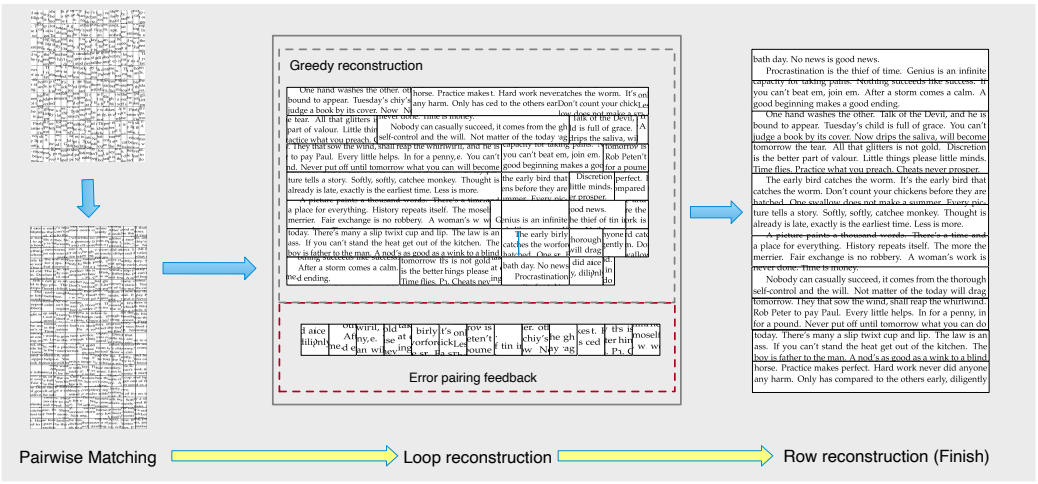


Figure 6: Schematic diagram of reconstruction process

pairs is small, it is easy to mark all the mismatched pairs with negative feedback. In each round of splicing, the splicing results are fed back through the GUI. After feedback, the correct splicing and the wrong splicing can be distinguished. When the matching fragments are true neighbors, the stitched neighbor fragments will not be matched by other fragments again, and can be removed from the candidate set, which results in a narrowing of the candidate set of mismatches or the replacement of matched fragments by new candidates. The entire splicing process is shown in Fig. 6.

3. Experiments

3.1. Test dataset and experimental environment

In order to eliminate the influence of some unpredictable factors on the feasibility of the algorithm and the accuracy of the splicing test, a feasible approach is to analyze the performance of the algorithm using a computer simulated dataset (Xu et al. (2014); Gong et al. (2016); Chen et al. (2019)).

In the experiment, we used an open source dataset[2] that included refactoring of six documents. In addition, a more complex and comprehensive dataset containing 70 test documents was created by digitally simulating the

[2]China Undergraduate Mathematical Contest in Modelling (2013) CUMCM-2013 contest problems

Table 1: Test dataset classification

Data number	Size of the picture	Scale	Dataset characteristics
D1 —— D3	72 × 1980px	1 × 19	Open source dataset of SSTD
D4 —— D7	72 × 180px	11 × 19	Open source dataset of CCSTD
D8 —— D19	204 × 330px	6 × 6	CCSTD in Chinese and English
D20 —— D31	136 × 220px	9 × 9	CCSTD in Chinese and English
D32 —— D43	68 × 330px	6 × 18	CCSTD in Chinese and English
D44 —— D55	68 × 220px	9 × 18	CCSTD in Chinese and English
D56 —— D67	68 × 180px	11 × 18	CCSTD in Chinese and English
D68 —— D72	72 × 180px	11 × 18	CCSTD containing images and text
D73 —— D78	72 × 180px	10 × 16	Image document of CCSTD

shredder cutting process. We noticed that the dataset used by Gong et al. (2016) contained a large number of blank fragments. In order to better reflect the performance of the algorithm and ensure the accuracy of experimental data, our dataset did not contain blank fragments. All of the algorithms were implemented in Python, and the tests were performed on a Core i7 6700k CPU with 8GB RAM. The characteristics of all datasets are summarized in Table 1:

3.2. Comparison and analysis of reconstruction accuracy

3.2.1. Evaluation method

In order to compare the accuracy of paired matches, a comprehensive evaluation strategy is required. There are two methods to evaluate the splicing progress of RCCSTD problems. The first method is based on the number of fragments in the wrong position, and the second method is based on the number of wrong splicing pairs. This paper uses the second method to evaluate the accuracy of reconstruction. In all paired matches, the wrongly matched pairs are counted, and the proportion of the number of wrong matches represents the splicing accuracy. In the experimental comparison, we defined the calculation formula of the splicing accuracy as follows:

$$accuracy = 1 - \frac{\sum_{i=1}^{n} p_i}{n}, \tag{10}$$

where n is the number of fragments; $p_i$is defined as follows:

$$\begin{cases} p_i = 1 & \text{if the ith scrap is matched to the error neighbor,} \\ p_i = 0 & \text{otherwise.} \end{cases} \tag{11}$$

3.2.2. Pairwise accuracy comparison

In the experiment, the two-step strategy algorithm Chen et al. (2019) and a splicing-driven memetic algorithm (SD-MA) Gong et al. (2016) are compared. In addition, the conventional algorithm sum of squared distances (SSD) was added as a comparison to analyze the effectiveness of the algorithm, and the accuracy of the proposed algorithm without IE preprocessing was used to evaluate the impact of the IE operation on pairing. Considering the lack of literature reference in RCCSTD problem of the same type of machine learning algorithm, the RCCSTD problem was defined as a binary classification problem, and BP neural network algorithms (Rumelhart et al. (1986)) were used to try to solve it. The datasets were grouped according to the scale of fragments in the dataset and the source of the dataset, and 20 datasets were randomly selected from each type of dataset in Table 1 for testing and comparison. Due to the instability of the evolutionary algorithm and the difference in the results of multiple runs, the process of solving the TSP is to run 20 times to get the best result. The visual display of the test results of these 20 datasets is shown in Table 2.

Table 2: Comparison of matching accuracy of several algorithms on datasets

Datasets	N	two-step strategy algorithm	SD-GA	SSD	BP	Proposed	Proposed without IE
D1	19	100.0% =	100.0% =	21.05% -	82.35% -	100.0%	100.0% =
D2	19	100.0% =	100.0% =	13.23% -	79.41% -	100.0%	100.0% =
D3	19	100.0% =	100.0% =	15.49% -	80.88% -	100.0%	100.0% =
D4	209	64.59% -	53.3% -	33.49% -	66.98% -	95.22%	87.55% -
D5	209	64.59% -	48.00% -	18.66% -	67.94% -	93.3%	89.5% -
D6	209	49.76% -	47.80% -	22.00% -	68.42% -	89.5%	88.5% -
D7	209	63.64% -	62.20% -	17.22% -	69.86% -	93.3%	89.5% -
D8	36	100% +	97.2% +	77.0% -	75.00% -	94.4%	88.9% -
D9	36	93.83% -	88.89% -	72.22% -	66.67% -	94.4%	88.9% -
D13	36	100.0% =	100.0% =	41.7% -	72.22% -	100%	100.0% =
D14	36	100% +	94.4% -	41.7% -	80.56% -	94.4%	97.2% +
D20	81	50.6% -	87.6% -	56.8% -	67.90% -	91.4%	87.7% -
D21	81	95.0% +	83.9% -	48.1% -	71.60% -	92.6%	79.0% -
D25	81	91.4% +	81.0% -	38.3% -	76.54% -	87.7%	84.0% -
D26	81	80.2% -	75.3% -	38.3% -	67.90% -	86.4%	85.1% -
D32	108	76.0% -	59.3% -	63.0% -	65.43% -	97.2%	94.4% -
D37	108	97.2% +	55.6% -	26.9% -	70.37% -	93.5%	92.6% -
D44	108	80.2% -	71.2% -	62.3% -	63.58% -	94.4%	91.3% -
D49	108	96.2% +	54.3% -	37.0% -	62.34% -	87.7%	85.2% -
D56	198	48.5% -	48.5% -	65.2% -	59.60% -	92.9%	90.4% -
D61	198	85.4% -	45.5% -	39.8% -	74.75% -	87.3%	82.8% -

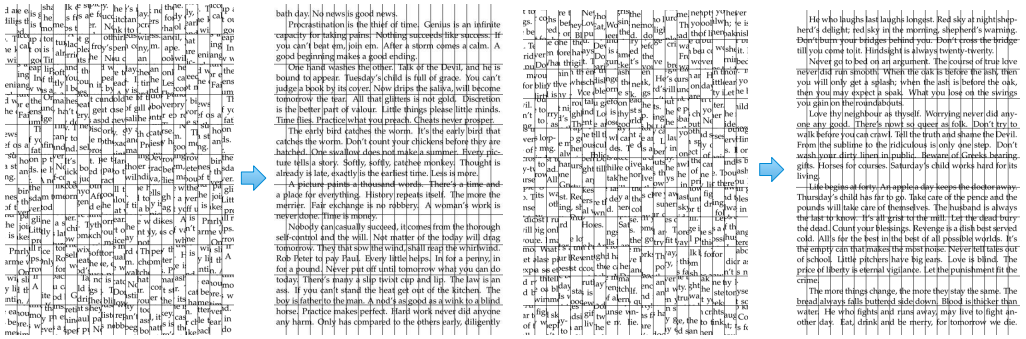
In order to understand the statistical difference between the algorithms,

the two-sample t-statistic test Keselman et al. (2004) was carried out to indicate significance between different results at the 0.05 significance level. As reported in Table 2, the significance of the test results is represented in a "/=/+" manner, where "+", "–", and "=" indicate that the results of the compared algorithms are significantly better than, worse than, or similar to that of the proposed algorithm, respectively. In addition, the best metric values are highlighted in gray.

The results show that the clustering algorithm can only ensure high accuracy when the number of text rows in the fragment is large. Furthemore, the matching effect is not ideal when the document is cut into smaller fragments. As for the SD-GA based on two-dimensional TSP search (Gong et al. (2016)), SD-GA can only ensure the recovery of large pieces of document, it is not even as good as conventional SSD algorithms on dataset D56. When the pieces of the document become smaller and more fragmented, the accumulation of splicing errors will increase the number of global errors and reduce the splicing accuracy rapidly. In the case of documents being cut into smaller pieces, our approach achieved satisfactory matching accuracy, especially on open source datasets D4 to D7. One of the reasons for this is that by preprocessing fragments and making full use of the limited pixel information of the fragments, errors that occurred in some special cases were corrected to some extent. As can be seen from Table 2, the pairing accuracy of the algorithm was improved after the use of the IE strategy. Another important element is the pairing-compatibility measurement model based on the extreme learning machine algorithm. A system with expert knowledge can be obtained by using consensus information of matched fragments for supervised learning, which can improve the matching accuracy while reducing the over-sensitivity to edge pixel. The original documents relating to datasets 5 and 6 are shown in Fig. 7.

3.3. Comprehensive dataset testing

RCCSTD based on clustering can only reconstruct special text documents, which are required to have obvious row features (Fig. 9). For document fragments without row feature pairs, the clustering accuracy will deteriorate rapidly. Our algorithm does not depend on a specific text format, so it can be applied to all types of documents. In our experiment, we added some special test sets, including a text with both text and a picture, and pure picture text. Satisfactory results were obtained. As shown in Table



(a) (b)

Figure 7: The reconstruction results for dataset 4 (a) and dataset 5 (b)

Table 3: Accuracy of paper fragments containing images

test data	D67	D69	D70	D71	D73	D75	D76
accuracy	91.4%	92.9%	95.9%	92.9%	98.1%	96.2%	95.6%

3, algorithm achieve high splicing accuracy on all complex datasets, indicating that the algorithm can be applied to various complex datasets. The reconstruction results for dataset 75 are shown in Fig. 8.

3.4. Analysis of candidate set strategy effectiveness

In order to improve the splicing efficiency and recover all the pieces of paper completely, we adopted the method of candidate neighbors in the semi-automatic splicing process, provided the negative feedback of wrong splicing manually, and conducted secondary screening to achieve the purpose of correct splicing. The effectiveness of the candidate set is reflected in the reduction of the number of manual interventions and improvement in the efficiency of auxiliary manual splicing. The expected result is that no manual intervention is needed, but as described in Section 1, it is extremely difficult to achieve automatic and accurate splicing for relatively fragmented text. The current algorithm can only improve the splicing accuracy to improve efficiency. The analysis of the number of corrections of the candidate sets is shown in Table 4. Even if the splicing had an error, with a limited amount of manual feedback, all fragments could be correctly matched, demonstrating the method is feasible and effective.

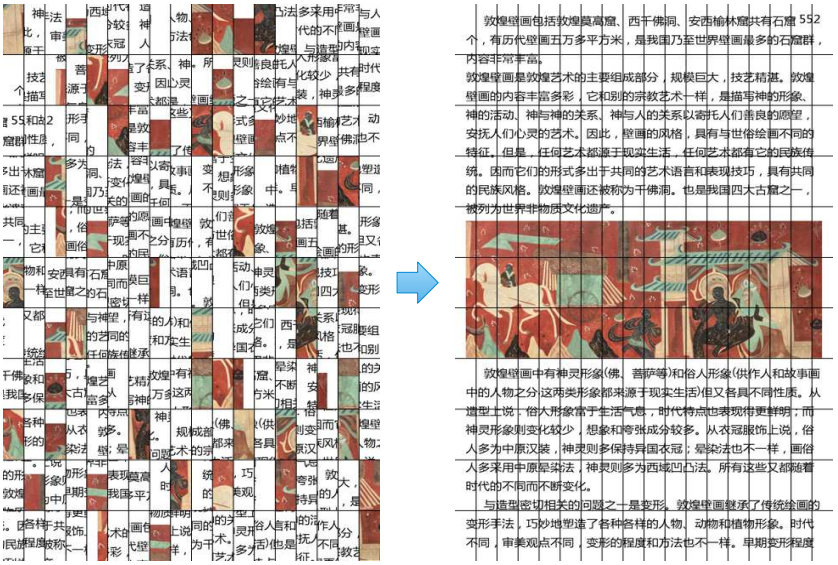


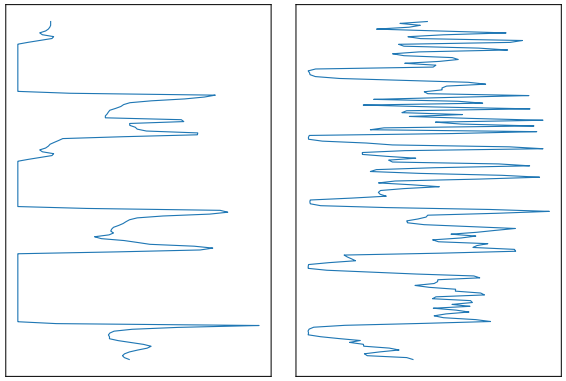
Figure 8: The result of a document containing text and images

3.5. Simulate noise in real environment

The simulation data represent, to a large extent, the ideal test environment. In practical problems, there may be many unpredictable factors, such as scanner accuracy or shredder cutting leading to irregular edges, resulting in uncontrollable interference. Even with image denoising, it is difficult to eliminate all external interference. Both the performance of the scanner and the image preprocessing algorithm may affect the accuracy of the splicing. The image preprocessing algorithm is not the focus of this paper, so we will not discuss it. According to the edge feature information of the shredded

Table 4: Influence of feedback on splicing accuracy

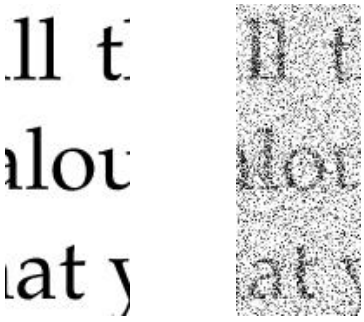
dataset	D3	D4	D5	D6
Accuracy without feedback	95.2%	93.3%	89.4%	93.3%
Mark the number of incorrect pairs	10	14	22	14
Accuracy after feedback	98.08%	100%	99.04%	99.5%
dataset	D36	D37	D48	D49
Accuracy without feedback	93.5%	92.5	87.6%	85.1%
Mark the number of incorrect pairs	7	8	20	24
Accuracy after feedback	100%	98.1%	98.1%	96.2%



(a) (b)

Figure 9: Edge information with row features (a) and edge information without edge features (b)

paper under real conditions, we find that Gaussian noise can be used to replace various complex situations in the actual scene, so we use the method of adding Gaussian noise to test the robustness of the algorithm (Fig. 10). Because the pixel information of the whole scrap paper is used, it is not feasible to simply add edge noise points. Gaussian noise points should be added to the whole picture.



(a) (b)

Figure 10: Comparison chart of text information with (a) and without noise (b).

In the experiment, the intensity of noise is controlled to simulate various complex scenes. For data with different interference levels, Gaussian noise with different variances is added to the picture in the experiment, and the

Table 5: Accuracy under different levels of noise

variance	0.001	0.002	0.004	0.006	0.008	0.01	0.1
D3	93.7%	92.8%	92.3%	88.9%	88.0%	85.6%	77.5%
D4	91.3%	90.4%	83.7%	85.1%	79.9%	81.3%	66.0%
D5	88.5%	88.9%	87.5%	86.6%	82.7%	80.8%	64.1%
D6	91.8%	90.9%	88.9%	84.2%	79.4%	76.0%	65.0%

useful pixel information for pairing also decreases as the noise increases. As shown in Table 5, if the noise cleaning of the image is not performed, the splicing accuracy decreases as the noise level increases. However, our algorithm still maintains high splicing accuracy. To some extent, this can reflect the strong anti-interference ability of the algorithm, which means that the algorithm can be used as an expert system in the actual production environment.

4. Conclusion and future work

In this paper, we have proposed a pairing compatibility measurement model based on the extreme learning machine algorithm and used the improved consensus information to obtain the pairing compatibility of the fragments in the RCCSTD problem. To completely reconstruct the document, all the pairing information was combined by greedy reconstruction, and error corrections were applied by manually marking incorrectly matched pairs. The complete restoration of all the fragments was realized with a limited amount of intervention. Satisfactory results were obtained in 78 RCCSTD instances with different languages and formats in the experiment.

In future work, we will determine how to obtain higher compatibility discrimination through neural networks. Although the method proposed in this paper achieved the best performance on all datasets, some datasets still needed manual intervention. By establishing a global relational network of fragments, and considering the adhesion and mutual exclusion between fragments, automatic correction of partial errors can be performed. In addition, the introduction of convolution kernels to extract the correct edge matching features is another future research direction. The method of capturing potential connections through consensus information between two objects in the RCCSTD problem can also be applied to other fields, such as video splicing, panoramic image synthesis, and informatics. The RCCSTD problem needs

to be further deepened to promote the use of the more robust and efficient RCCSTD expert system in actual scenarios and scenario expansion.

Acknowledgement

The authors wish to thank the support of the National Natural Science Foundation of China (Grant No. 61876164, 61673331, 61772178), the Education Department Major Project of Hunan Province(Grant No. 17A212), The MOEA Key Laboratory of Intelligent Computing and Information Processing, the Science and Technology Plan Project of Hunan Province (Grant No. 2016TP1020), the Provinces and Cities Joint Foundation Project (Grant No. 2017JJ4001), the Hunan province science and technology project funds(2018 TP1036).

References

Biesinger, B., Schauer, C., Hu, B., & Raidl, G. R. (). Reconstructing cross cut shredded documents with a genetic algorithm with solution archive. Citeseer.

Chen, J., Tian, M., Qi, X., Wang, W., & Liu, Y. (2019). A solution to reconstruct cross-cut shredded text documents based on constrained seed k-means algorithm and ant colony algorithm. *Expert Systems with Applications*, .

Dekel, T., Oron, S., Rubinstein, M., Avidan, S., & Freeman, W. T. (2015). Best-buddies similarity for robust template matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2021–2029).

Freeman, H., & Garder, L. (1964). Apictorial jigsaw puzzles: The computer solution of a problem in pattern recognition. *IEEE Transactions on Electronic Computers*, (pp. 118–127).

Gong, Y.-J., Ge, Y.-F., Li, J.-J., Zhang, J., & Ip, W. (2016). A splicing-driven memetic algorithm for reconstructing cross-cut shredded text documents. *Applied Soft Computing*, *45*, 163–172.

Heingartner, D. (2003). Back together again. *New York Times*, *17*.

Huang, G.-B., & Babri, H. A. (1998). Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions. *IEEE transactions on neural networks*, *9*, 224–229.

Huang, G.-B., Zhu, Q.-Y., Siew, C.-K. et al. (2004). Extreme learning machine: a new learning scheme of feedforward neural networks. *Neural networks*, *2*, 985–990.

Jain, A., Nandakumar, K., & Ross, A. (2005). Score normalization in multimodal biometric systems. *Pattern recognition*, *38*, 2270–2285.

Jain, A. K., Mao, J., & Mohiuddin, K. (1996). Artificial neural networks: A tutorial. *Computer*, (pp. 31–44).

Keselman, H., Othman, A. R., Wilcox, R. R., & Fradette, K. (2004). The new and improved two-sample t test. *Psychological Science*, *15*, 47–51.

Li, R., Zhang, H., Gao, S., Wu, Z., & Guo, C. (2021). An improved extreme learning machine algorithm for transient electromagnetic nonlinear inversion. *Computers & Geosciences*, *156*, 104877.

Lin, H.-Y., & Fan-Chiang, W.-C. (2012). Reconstruction of shredded document based on image feature matching. *Expert Systems with Applications*, *39*, 3324–3332.

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly weather review*, *116*, 2417–2424.

Paikin, G., & Tal, A. (2015). Solving multiple square jigsaw puzzles with missing pieces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4832–4839).

Paixão, T. M., Boeres, M. C., Freitas, C. O., & Oliveira-Santos, T. (2019). Exploring character shapes for unsupervised reconstruction of strip-shredded text documents. *IEEE Transactions on Information Forensics and Security*, *14*, 1744–1754.

Prandtstetter, M. (2009). *Hybrid optimization methods for warehouse logistics and the reconstruction of destroyed paper documents*. na.

Prandtstetter, M., & Raidl, G. R. (2008). Combining forces to reconstruct strip shredded text documents. In *International Workshop on Hybrid Metaheuristics* (pp. 175–189). Springer.

Ranca, R. (2013). A modular framework for the automatic reconstruction of shredded documents. In *Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*, 533–536.

Sleit, A., Massad, Y., & Musaddaq, M. (2013). An alternative clustering approach for reconstructing cross cut shredded text documents. *Telecommunication Systems*, *52*, 1491–1501.

Ukovich, A., Ramponi, G., Doulaverakis, H., Kompatsiaris, Y., & Strintzis, M. (2004). Shredded document reconstruction using mpeg-7 standard descriptors. In *Proceedings of the Fourth IEEE International Symposium on Signal Processing and Information Technology, 2004.* (pp. 334–337). IEEE.

Wang, Y., & Ji, D.-C. (2014). A two-stage approach for reconstruction of cross-cut shredded text documents. In *2014 Tenth International Conference on Computational Intelligence and Security* (pp. 12–16). IEEE.

Xu, H., Zheng, J., Zhuang, Z., & Fan, S. (2014). A solution to reconstruct cross-cut shredded text documents based on character recognition and genetic algorithm. In *Abstract and Applied Analysis*. Hindawi volume 2014.

Zhu, L., Zhou, Z., & Hu, D. (2007). Globally consistent reconstruction of ripped-up documents. *IEEE Transactions on pattern analysis and machine intelligence*, *30*, 1–13.